Coping With Imbalanced Data in the Automated Detection of Reminiscence From Everyday Life Conversations of Older Adults

Stoev, Teodor ; Ferrario, Andrea ; Demiray, Burcu ; Luo, Minxia ; Martin, Mike ; Yordanova, Kristina

DOI: https://doi.org/10.1109/access.2021.3106249

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 14, 2021, accepted August 13, 2021, date of publication August 18, 2021, date of current version August 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3106249

Coping With Imbalanced Data in the Automated Detection of Reminiscence From Everyday Life Conversations of Older Adults

TEODOR STOEV[1], ANDREA FERRARIO[2], BURCU DEMIRAY[3], MINXIA LUO[3], MIKE MARTIN[4], AND KRISTINA YORDANOVA[1]

[1]Institute for Visual and Analytic Computing, University of Rostock, 18059 Rostock, Germany
[2]Mobiliar Lab for Analytics at ETH, ETH Zurich, 8092 Zurich, Switzerland
[3]Department of Psychology, University of Zurich, 8006 Zurich, Switzerland
[4]Gerontology Center, University Research Priority Program "Dynamics of Healthy Aging," University of Zurich, 8006 Zurich, Switzerland

Corresponding author: Teodor Stoev (teodor.stoev@uni-rostock.de)

The work of Teodor Stoev and Kristina Yordanova was supported in part by the German Research Foundation (DFG), grant number YO 226/3-1, and in part by the University of Rostock within the funding program Open Access Publishing.

ABSTRACT Reminiscence—the act of recalling or telling others about relevant personal past experiences—plays an important role in the well-being of older adults. Therefore, it is relevant to develop intelligent systems aiming at improving the well-being of the elderly by reliably detecting reminiscence in their everyday life conversations. Data imbalance is one of the main challenges in the automatic detection of reminiscence from everyday conversations, as reminiscing is a rare event. In this paper, we address the problem by proposing a methodology for coping with imbalanced data in the detection of reminiscence in conversations of older adults. The methodology combines data augmentation using BERT (Bidirectional Encoder Representations from Transformer) and feature extraction techniques leveraging natural language processing for the German language. We evaluate the proposed methodology on a dataset comprising transcripts of social conversations of older adults held in German. We compare our results with a previous work addressing the problem on the same dataset and we show that our approach strongly outperforms the baseline. The results in this study may support the development of intelligent systems for the real-time detection of reminiscence in everyday life of older adults and the design of digital health interventions to support their well-being.

INDEX TERMS Natural language processing, data augmentation, BERT, machine learning, reminiscence, well-being, older adults.

I. INTRODUCTION AND MOTIVATION

Reminiscence, i.e., the "naturally occurring act of thinking about or telling others about personally meaningful past experiences" [1], [2], is a crucial activity in aging individuals [3]. The experiences recalled through reminiscing can vary from specific single events, to repeating occurrences of an event, and experiences covering long periods of time. Sharing and recalling past experiences can support decision making, self-understanding, or bonding with others [4], [5]. Reminiscing can also improve the well-being of people and help them feel social inclusion in a period of their life when they are no longer actively working, or participating in many of the social functions of our society [4]. Sharing reminiscence events with others is commonly referred to as "social reminiscence" [1].

Reminiscence is an object of study also in the intelligent systems computing research domain. In fact, reminiscence is a "context-dependent, real-world cognitive activity" [6], [7]: if it can be reliably and continuously detected, then it is possible to study its effects on mental and health outcomes of older adults in their everyday life. Therefore, studies considered the elicitation of reminiscence events with emails, random text messages, or videos, music and photographs [8]–[10], through the use of tangible objects, such as chests of drawers coupled with digital interfaces [11], or the representations of digital object memories [12]. Others investigated the collection of personal data by ubiquitous tools to assist users

The associate editor coordinating the review of this manuscript and approving it for publication was Ziyan Wu.

in self-reflection, including personal informatics activity and reminiscing [13].

As opposed to the elicitation of reminiscence with various technologies or by retrospective self-reports [13], [14], the real-time detection of social reminiscence (or the lack thereof) in everyday settings could play a crucial role in designing intelligent systems aiming at supporting the well-being of older people by means of unobtrusive and reliable technology [6]. These well-being supporting systems (comprising, for example, smartphone apps) may combine early warning algorithms that detect reminiscence events from everyday conversations of older adults in real-time, classifying them into positive and negative functions [15] and triggering digital health intervention programs [16] aimed at coping with the cases of negative reminiscence. These latter, such as the act of ruminating on the past or maintaining intimacy (i.e., feeling close to people who are no longer part of one's life), have been proven to result in a detriment of physical and mental well-being [17], [18].

However, the deployment of these intelligent systems, together with the design of digital health intervention programs focusing on the activity of reminiscing, are still fraught with a few challenges. One of the main challenges of detecting reminiscence in social conversations is represented by the imbalance between reminiscence events and all other types of conversations. In fact, reminiscence in everyday life is a rather rare event; studies have shown that it comprises only about 5% of all everyday conversations of older adults [1], [6], [19].

Another challenge is represented by the need to distinguish between reminiscence and all the past tense conversations that are not related to personal past experiences. Data generation procedures may exacerbate this challenge, as in multiple studies the app used to record the conversations of older adults stored only 30-seconds long audio snippets, due to privacy reasons [1], [6], [20]. This resulted in a quite diverse corpus of recorded conversations (or parts of), some including full sentences and periods, some just few interjections [6]. Finally, previous studies where conversations of older adults are recorded and subsequently transcribed, comprise only few thousands of transcripts, affecting the performance of machine learning classifiers [6], [21].

In a previous work, Ferrario *et al.* conducted a feasibility study that investigated the applicability of machine learning and natural language processing (NLP) methods to the problem of automatic detection of social reminiscence from daily conversations of older adults in German [6]. The results of that study show that it is possible to detect reminiscence by 1) preprocessing the transcripts of the conversations with NLP, and 2) training classifiers on different features by implementing learning strategies to cope with class imbalance. However, the aforementioned challenges affected, in particular, the precision of the best performing classifiers and lead to moderate values of recall (such as 0.45, for the best classifier in [6]).

The use of unobtrusive technology for the automated detection of reminiscence in older adults is of relevance for the domain of life span and aging research, as it may pay the way to the design of personalized digital health interventions. In fact, by the reliable and real-time monitoring of reminiscence in their everyday life, older adults may support their autonomy by observing, measuring, monitoring and acting for their own well-being, as recommended by the WHO healthy aging model [22].

This work uses the dataset of transcriptions of everyday conversations of older adults collected in Demiray *et al.*'s study [1]. The same dataset was used by Yordanova *et al.* to detect social behaviors and environments [21] and by Ferrario *et al.* [6] to detect reminiscence using NLP and machine learning. Based on the challenges identified in Ferrario *et al.*' s study, in this work we propose a methodology for the detection of reminiscence from textual data in the case where

1) the dataset is strongly imbalanced;
2) the texts contain snippets of conversations as opposed to full conversations;
3) the texts are noisy as they come from transcripts of short audio recordings where some of the words and context may be lost during transcribing;
4) and the language is German, which poses additional challenge as there are less language resources as compared to the English language.

The methodology we propose consists of three modules: data augmentation, feature extraction and binary classification (reminiscence vs. not reminiscence) with machine learning. In this work, our focus is on the first two modules of the proposed methodology. Furthermore, we investigate the impact of the augmentation factor and of different feature combinations generated from the textual data on the performance of the machine learning classifier.

The contributions of the paper are as follows:

1) we propose a methodology for the detection of reminiscence from the transcripts of daily conversations of older adults based on data augmentation using BERT (Bidirectional Encoder Representations from Transformer);
2) we investigate the impact of the data augmentation factor on the performance of the system when augmented data is introduced;
3) we investigate the impact of the different features on the performance of the machine learning classifiers and we propose a best set of features;
4) we show that the proposed methodology is able to detect reminiscence reliably in transcripts in the German language and to outperform the benchmark represented by Ferrario *et al.*'s best results [6].

The paper is structured as follows. In Section II we present the state of the art in the field of detection of social reminiscence and the more general field of automated coding in social sciences. Then, we look into text augmentation and feature extraction methods, with a focus on the German

language. Section III describes the proposed approach for the detection of reminiscence in social conversations in some detail, including the selected methods. Section IV describes the experimental setup for our evaluation. Section V presents the evaluation results and, finally, Section VI concludes the paper with a discussion on limitations of the current study and on the future work.

II. RELATED WORK

Automated Detection of Reminiscence From Everyday Conversations: The ability of reminiscing in older adults has been traditionally studied by means of self-reporting and life reviews [3] and automated reminiscence therapy [23]. In the former case, reminiscing is averaged throughout a given period of time. In the latter case, reminiscing is elicited from study participants by therapists or by using remote support. On the other hand, Demiray *et al.* [1] recently studied social reminiscence in a naturalistic observation study, i.e., by passively recording and transcribing everyday conversations of older adults for a period of four days. In particular, their study quantified the reminiscence activity (5% of all conversations) of participants and described the identified functions. The study provided a first analysis of reminiscence as ''naturally'' occurring activity in everyday life and in absence of direct elicitation by therapists or any remote support [1]. The participants in the study reminisced mostly with friends, partners and children or relatives. Reminiscing served different functions, such as identity, i.e., the sense of being the same person over time and the preservation of a positive self-concept, and teaching/informing others. Finally, Demiray *et al.* showed that reminiscing is positively related to life satisfaction and negatively with mood [1].

Using the dataset from the Demiray *et al.* study [1], Ferrario *et al.* proposed a methodology for the automated detection of social reminiscence from the transcripts of conversations in German of older adults [6]. The automated detection system made use of natural language processing (NLP) and machine learning classifiers. In particular, their methodology comprised the generation of different NLP features (bag-of-words, bag-of-part-of-speech tags, and pretrained German word embeddings) from the corpus of transcripts, and the use of different learning strategies with four families of classifiers (random forests, adaptive and extreme gradient boosting, support vector machines). The different learning strategies aimed at coping with the class imbalance [24], [25] in the corpus of transcripts. One of the learning strategies made use of data augmentation using Synthetic Minority Oversampling Technique (SMOTE) [26]. Their results showed that class-weighted support vector machines (SVM) on bag-of-words features outperformed all other classifiers (F1=0.48, precision=0.5, recall=0.45), followed by SVM on SMOTE-augmented data and pretrained word embeddings features (F1=0.44, precision=0.35, recall=0.59). All models in the Ferrario *et al.*'s study showed rather low precision and moderate values of recall, as shown, for example, by the best performing classifiers. All details are contained in Table 2, [6]. In fact, an error analysis showed that ''the models tend to predict long transcripts with multiple sentences referring to the past incorrectly as reminiscence'' [6]. This phenomenon, together with class imbalance in the provided dataset, affected the overall performance (i.e., the F1-score) of the classifiers and, *de facto*, limiting their applicability in intelligent systems to be used in the everyday life of older adults [6].

Automated Methods for Coding of Social Conversations: The automated detection of reminiscence makes use of reliable ground truth labels [6]. These are time-consuming to generate, as they involve the manual coding of transcripts by trained resources [1], [6]. Manual coding of ground truth labels for machine learning problems is a common process in social sciences [27]; a more recent and viable alternative is represented by automated coding. Yordanova *et al.* [21] introduced a natural language processing and machine learning-based approach for the automated coding of social behaviours and environments using transcripts of everyday conversations, including the data from Demiray *et al.* [1]. Their results showed that it is possible to automatically code social behaviour and environments with machine learning algorithms. In particular, they showed that the random forests algorithm [28] consistently outperformed decision trees and SVMs in all the proposed experiments [21]. Based on the promising results in [21], in our work we also use random forest as a classifier.

In another work, Crowston *et al.* [29] developed machine learning and natural language processing-based techniques to automate coding of textual data. They applied bag-of-words models and part-of-speech tags to generate NLP features. Their results showed that both rule-based and machine learning and natural language processing-based automatic coding can support coding of textual data. However, the latter automatic coding technique needs a large number of samples and suffers from unbalanced data distributions.

Methods to Augment Textual Data: Data augmentation techniques are commonly used in natural language processing tasks to cope with class imbalance in textual data. A widely used technique is SMOTE, or Synthetic Minority Oversampling Technique [26]. The SMOTE algorithm generates synthetic samples from the minority class to be augmented, by finding the K nearest neighbors of a given data point, and connecting them by line segments. As previously remarked, Ferrario *et al.* [6] used SMOTE as data augmentation technique to cope with class imbalance. However, in natural language processing applications, the main limitation of SMOTE is the fact that it does not take into consideration the structural properties of textual data. In fact, the SMOTE algorithm takes as input the numerical vector representation of textual data [26]. On the other hand, EDA (Easy Data Augmentation) techniques perform data augmentation by means of operations on the words contained in the textual data [30]. EDA techniques comprise 1) synonym replacement, 2) random insertion (of synonyms),

3) random swap, and 4) random deletion of words in textual data. EDA techniques have lead to the improvement of performance for both convolutional and recurrent neural networks on text classification tasks [30]. The use of synonym replacement has been previously explored by other authors, as well [31]. As noted by Yordanova *et al.* [21], textual data augmentation using synonyms can be applied only to a percentage of the vocabulary of the corpus at hand [32]. Yordanova *et al.* [21] used both synonym and hyperonyms replacement leveraging the taxonomy of English language WordNet [33] to augment textual data and cope with severe class imbalance. Hyperonyms introduce ''a 'type-of' relation with the original word'' [21]; they are considered whenever synonyms of a given word are not found.

One of the main limitations of synonym replacement is represented by the necessity to find a suitable source of words with similar meaning. This is done by utilizing lists of known synonyms or lexical resources such as WordNet [33]. However, these resources are not available for all languages and they are not applicable in contexts where a specific domain has to be handled. Studies showed that using word embedding models, such as word2vec [34], can be competitive for synonym replacement tasks by leveraging the previously mentioned linguistic resources [35]. On the other hand, next to synonym replacement and word embeddings, recent studies use context-based language models to perform textual data augmentation [32], [36], [37]. One example of such a model is BERT (Bidirectional Encoder Representations from Transformers) [38]. BERT is a neural network language model that takes into consideration the left and right context of a given word. In that way, BERT is trained by masking words from a sentence and allows predicting their empirical probability given the surrounding words [38]. A pretrained BERT model can be fine-tuned on downstream tasks by just adding one additional layer to the model. This makes BERT easy to use for many use cases such as text classification, question answering and language inference [38]. In our work we also rely on BERT to find synonyms for German language.

Finally, back-translation, i.e., the automated translation of a (mono)lingual target text into the source language is another textual data augmentation method. Sennrich *et al.* [39] used it in the WMT 15 English<–> German back-translation task. They showed that augmenting the training data set with back-translated samples delivers state-of-the-art results.

Feature Extraction From Textual Data of Daily Conversations: In previous studies different sets of features have been extracted from textual data to analyse written transcripts of daily conversations, to detect reminiscence, everyday social behaviours and environments [6], [21].

In particular, Ferrario *et al.* focused on the feature extraction from transcripts in the German language to detect reminiscence. Their results showed that pretrained word embeddings outperformed bag-of-words and (bag-of-) part-of-speech tags when data augmentation with SMOTE is used [6]. On the other hand, bag-of-words features outperform (bag-of-) part-of-speech tags and pretrained word embeddings when using class weighted learning [6]. Moreover, it is worthy to note that Ferrario *et al.* did not consider the combination of different families of features to detect reminiscence from the given corpus of transcripts [6].

Yordanova *et al.* considered three distinct datasets of transcripts both in English and German [21]. They performed Latent Dirichlet Allocation (LDA) to compute topics for the words in the corpora of transcripts. In addition, they considered part-of-speech tags and features obtained by computing the semantic similarities of words at a given level of abstraction with WordNet [21]. Finally, Yordanova *et al.* made use of time-dependent features, such as the day of the week, whether the day of the transcript is a weekend day or not, and the part of the day (morning, noon, afternoon, evening or night) [21]. In our work we also look into time-dependent features.

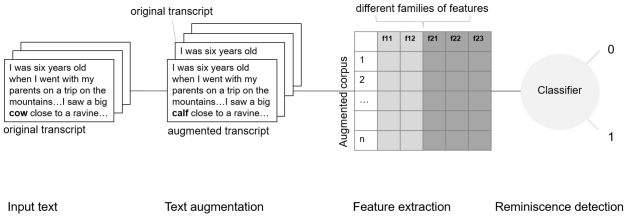


FIGURE 1. The proposed workflow for detection of reminiscences from transcripts of daily conversations. We show an example of transcript in English for the sake of readability.

III. PROPOSED APPROACH

In this section, we describe the proposed approach for the detection of reminiscence from a corpus of transcripts of everyday life conversations of older adults. The approach consists of three modules (see Figure 1). First, the input text (i.e., the corpus of transcripts of social conversations) is processed generating new data samples (**text augmentation**). Second, features are computed from the augmented corpus of transcripts (**feature extraction**). Lastly, the extracted features are used to train machine learning classifiers to detect reminiscence (**reminiscence detection**). In the remainder of this section, we discuss the three modules in some detail.

A. TEXT AUGMENTATION

In this study, we performed textual data augmentation on the corpus of transcripts of social conversations using the BERT (Bidirectional Encoder Representations from Transformer) language model [38].

First of all, we tokenized each textual data point t and we retrieved a random sample $s(t)$ comprising 40% of all the tokens in t. For each token $w \in s(t)$, we performed synonym replacement by 1) computing with BERT the distribution of probabilities of words in the context of w, and 2) randomly selecting one of the words in the top-100 rank of the probability distribution. We performed the above steps only for tokens which are not stopwords, and we applied the

TABLE 1. The abbreviation of all the features extracted from the corpus of augmented textual data and their description.

Abbreviation	Feature Description
BOW	Bag-of-words features
POS	Bag-of-POS features
EMB	Word embeddings
LOC, **ORG** and **PER**	Number of words identified as a location, an organization, and a person
AMT_PT	Number of words expressing past tense
PTL_4	Number of words expressing past tense in sentences of length greater or equal to four
TEMP_WORDS	Number of selected temporal words
DEP_AUX_PP	Number of dependencies "auxiliary verb and a verb in perfect tense"
TIME	Time of the day features
AMT_ICH	Amount of personal pronouns "ich"

synonym replacement only on textual data with a minimal length of four tokens.

B. FEATURE EXTRACTION

We computed the following NLP features from the corpus of augmented textual data: 1) "bag-of-" models on both words and part-of-speech (POS) tags, 2) real-valued embeddings using pretrained German word embeddings, 3) named entities (locations, organizations and persons), 4) number of words expressing past tense, 5) number of words expressing past tense in sentences of length greater or equal to four words, 6) number of selected temporal words, 7) number of "auxiliary-verb and past participle verb" dependencies, 8) time-of-day features and 9) amount of "ich" (i.e., "I") pronouns. We collect them in Table 1 for the sake of readability. In what follows, we provide a high-level description of all the extracted features and of the process leading to their generation.

1) "BAG-OF-" MODELS

(**BOW**, **POS**) "Bag-of-" models compute numerical representations of transcripts by tokenizing them, and counting the occurrences of unique tokens (i.e., "bag-of-word" models **BOW**), and the part-of-speech (POS) tags (i.e., "bag-of-POS" models **POS**) collected from them. Afterwards, term frequency-inverse document frequency (tf-idf) normalization can be applied to the counts, for all "bag-of-" models. As a result, "bag-of-" models produce a real-valued matrix $B \in \mathbb{R}^{n \times m}$, where n denotes the number of available transcripts and m is the number of unique words, respectively POS tags in the given corpus.

2) WORD EMBEDDINGS

(**EMB**) Word embeddings are real-valued representations of textual data, such as the written transcripts of conversations of elderly adults, which make use of the "distribution hypothesis" by Harris [40]. The hypothesis states that words occurring in similar contexts should have "close" real-valued representations. The word embedding representations are numerical vectors of fixed dimension that depends on the word embedding model at hand. Embeddings can be trained on selected corpora of textual data, or pretrained embeddings can be leveraged. These latter may be suitable in presence of few data points. In this study, we considered pretrained word embeddings, only.

3) NAMED ENTITIES

(**LOC**, **PER**, **ORG**) Named entities are words that indicate a location (**LOC**), an organization (**ORG**), and a person (**PER**) [41]. We extracted named entities to improve the characterization of the context of the conversations contained in each transcript of the given corpus, as the number and type of named entities in a transcript may support the identification of narrative speech, such as in the case of reminiscence.

4) NUMBER OF WORDS EXPRESSING PAST TENSE

(**AMT_PT**) As reminiscing is the act of recalling personal past experiences, we counted the number of words with verb subcategory POS-tags "VVPP," "VAPP" and "VMPP." These tags are encoded in the TIGER Treebank annotation scheme [42] and they correspond to the participle forms "Partizip Perfekt, voll," "Partizip Perfekt, aux" and "Partizip Perfekt, modal" that are commonly used in the German language when talking about past events.

5) NUMBER OF WORDS EXPRESSING PAST TENSE IN SENTENCES OF LENGTH GREATER OR EQUAL TO FOUR

(**PTL_4**) We also counted the number of words expressing past tense in sentences of length greater or equal to four using the same methodology from Section III-B4. We considered the length of a sentence to discriminate between sentences referring to past events and those that simply contain a past participle form of a verb. We considered the length of four tokens as in the German language three words are needed to form a sentence in the past participle form (i.e., the subject, an auxiliary verb and the past participle of a verb).

6) NUMBER OF SELECTED TEMPORAL ADVERBS

(**TEMP_WORDS**) We also counted the number of words that are typically used in the German language to express an action or a situation in the near past. The list of words we considered comprises "jetzt," "heute," "zurzeit," "eben," "gerade," "gestern," "vorgestern," "momentan," "aktuell" (i.e., "now," "today," "currently," "just," "just," "yesterday," "the day before yesterday," "at the moment," "currently").

7) NUMBER OF DEPENDENCIES "AUXILIARY-VERB AND PAST PARTICIPLE VERB"

(**DEP_AUX_PP**) For each transcript, we counted the number of dependencies which contain an auxiliary verb and a past participle verb. We considered this dependency as in the German language it encodes all the contexts when a person is talking about the past.

8) TIME OF THE DAY FEATURES

(**TIME**) For each transcript in the dataset of this study, we computed the time of the day of the audio file corresponding to the given transcript. To do so, we partitioned a day into five consecutive time intervals and coded them as follows: from 6am to 11am (''morning''), from 11am to 13am (''noon''), 13am to 17pm (''afternoon''), from 17pm until 22pm (''evening'') and from 22pm until 6am (''night''). These features were inspired by a previous work by Yordanova *et al.* [21] where they were used as the authors assumed that depending on the time of the day people express different social behaviours (including reminiscing).

9) AMOUNT OF "ich" (i.e., "I") PRONOUNS

(**AMT_ICH**) For each transcript in the given corpus, we counted the number of ''ich'' (i.e., ''I'') pronouns. We considered this feature, as it helps encoding contexts where people talk about the personal past, i.e., when they talk about themselves and the situations that they experienced.

C. REMINISCENCE DETECTION

We performed the automated detection of reminiscence by training random forest classifiers on the sets of features extracted from the corpus of augmented transcripts as in Section III-B. Random forest classifiers have been used by Yordanova *et al.* for the automated coding of social behaviours and environments [21] on datasets including the one of this study, and by Ferrario *et al.* [6] for the automated detection of reminiscence. Random forests [28], [43] are machine learning algorithms that, in the case of classification problems, construct an ensemble of decision trees and generate predictions computing the mode of the classes returned by each tree. As in this study we focused on data augmentation and feature generation, we selected only one class of machine learning algorithms (i.e., random forests) to detect reminiscence from the given corpus of transcripts.

IV. EXPERIMENTAL SETUP

A. RESEARCH HYPOTHESES

Motivated by the literature on the use of different feature combinations in tasks with transcripts of social conversations [6], [21], [30], and on textual data augmentation (e.g., by synonym replacement) [30], in this study we introduced the following hypotheses:

- **H1**: different combinations of features extracted from textual data can lead to performance improvement with respect to the use of a single type of features in the task of detecting reminiscence.
- **H2**: data augmentation can lead to a performance improvement with respect to non-augmented data in the task of detecting reminiscence.

To test **H1**, we performed experiments with no data augmentation and by training the random forest classifiers on combinations of the types of features from Section III-B. On the other hand, to test **H2** we performed experiments with different augmentation factors and by training the random forest classifiers on the types of features (and their combinations) from Section III-B. All results are discussed in Section V.

B. DATASET

The dataset considered in this study comprises manually coded transcripts of everyday conversations in German of older adults. Data are originally collected in the naturalistic observation study by Demiray *et al.* [1]. In that study, 48 older adults (22 men and 26 women of 62–82 years of age) residing in Zurich, Switzerland, wore the Electronically Activated Recorder [20] over a period of 4 days. The EAR randomly recorded ambient sound (72 snippets of 30 seconds per participant per day), unobtrusively. Therefore, each participant was recorded 288 times throughout the experiment and for a total 144 minutes. The study generated 13,275 audio files, out of which 2214 (17%) contained conversations in Swiss German of older adults. The Swiss German dialect was translated word-by-word into standard written German [44]. Each conversation was then manually transcribed and manually coded by two independent coders, for different social behaviour and environment codes (see [21] for more details), as well as the function of reminiscence [1], [6]. Out of 2214 transcripts of conversations, 109 (4.9%) were coded as reminiscence.

C. PROCEDURE

In this section, we provide details on the procedure followed in our study, by describing how we implemented all the modules from Section III (see Figure 1) and specifying their resources. We close this section with a discussion on the labels of augmented transcripts.

1) IMPLEMENTATION OF THE **TEXT AUGMENTATION** MODULE

We implemented the **text augmentation** module described in Section III, as follows. As in Ferrario *et al.*'s study [6] we performed a stratified random split of the corpus of transcripts into training and test datasets, with an 80:20 ratio. The train dataset comprises 1771 samples (87 reminiscence) and the test dataset contains 443 samples (22 reminiscence). We then augmented the reminiscence samples in the training dataset by choosing an augmentation factor f, where $f \in \{0, 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$. The factor f is the number of artificially generated samples from each reminiscence sample. Therefore, the case $f = 0$ corresponds to no data augmentation. To improve the robustness of the machine learning results, we repeated the above procedure, i.e., the randomly stratified split of the corpus

of transcripts into training and test datasets followed by data augmentation, five times by choosing five distinct random seeds. In the **text augmentation** module we used the contextual word augmentation functionality provided in the Python library `nlpaug` [45] and the BERT model `bert-base-german-cased-oldvocab` available at `www.huggingface.co`.

2) IMPLEMENTATION OF THE **FEATURE EXTRACTION** MODULE

In the **feature extraction** module, we used the Python library `scikit-learn` to generate bag-of-words models and to apply tf-idf normalization. Similarly to Ferrario *et al.*, we generated POS tags using the POS tagger provided in the ''de_core_news_sm'' model for the German language in the `spacy` Python library [6]. It comprises 55 distinct tags. The same `spacy` model was used by Ferrario *et al.* to compute pretrained word embeddings [6]. We used it in this study to compute a 300-dimensional representation of each transcript, by averaging the embeddings of the tokens generated from the transcript after its tokenization. The features time of the day (**TIME**) and amount of ''ich'' (i.e., ''I'') pronouns (**AMT_ICH**) have been extracted using the transcripts metadata and by string matching, respectively. We used the `spacy` NER German model `de_core_news_lg` to extract the remaining features presented in this study.

3) IMPLEMENTATION OF THE **REMINISCENCE DETECTION** MODULE

We used the random forest implementation in the Python library `scikit-learn` as machine learning algorithm to detect reminiscence [46]. We chose an ensemble of 300 trees, with maximal depth equal to 15. The choice of the number of trees in the ensemble and their depth is motivated by a series of preliminary experiments on the dataset of transcriptions where they outperformed alternative choices of hyper-parameters (using grid search). We measured the performance of the random forest classifiers on the different combinations of the features in Section III-B by computing the precision, recall and F1-score, i.e., the harmonic mean of precision and recall, on the test dataset, for all five repeats. Finally, we reported the mean precision, recall and F1-score on the five repeats, for all trained random forest classifiers.

4) DOES DATA AUGMENTATION CONSERVE TRUE LABELS?

Performing data augmentation one generates new samples by modifying existing ones (i.e., transcripts) while their class labels are maintained, instead. However, as originally noted by Wei and Zou ''[i]f sentences are significantly changed, however, then original class labels may no longer be valid'' [30]. Therefore, it becomes relevant to check whether the class labels of the new samples are those of the original data samples. To do so–and differently from [30], where recurrent neural networks (RNNs) are implemented–we use the `spacy` pretrained embeddings [47] on the corpus of augmented data and we visualize them in 2D using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [48], [49]. t-SNE performs dimensionality reduction (into two or three dimensions) by embedding data point in such a way that similar data points are nearby and dissimilar data points result far away from each other, in a probabilistic sense [48], [49].

TABLE 2. Top performing features when training the random forest classifier with no data augmentation.

Features	Precision	Recall	F1
AMT_PT	0.226	0.045	0.073
POS	0.400	0.018	0.035
DEP_AUX_PP	0.300	0.018	0.034
PTL_4	0.267	0.018	0.033

As an example, in Figure 2 we show the results of the procedure in presence of the augmentation factor equal to f=20. The samples resulting from data augmentation closely surround the corresponding original data points. This suggests that the new data points conserve the labels of the corresponding original ones.

D. BASELINE

Ferrario *et al.* [6] used the dataset of 2214 transcripts from Demiray *et al.*'s study [1] for the automated detection of reminiscence. Their approach consisted in 1) the generation of NLP features (bag-of-words models, part-of-speech tags, pretrained German word embeddings), and 2) the training of machine learning models (random forests, support vector machines, and adaptive and extreme gradient boosting algorithms) in three distinct learning strategies. These strategies are introduced to cope with class imbalance in the data, by 1) class-weighted learning, 2) learning of a meta-classifier consisting of a voting ensemble, and 3) data augmentation with SMOTE. Their results showed that support vector machines on 1) bag-of-words features for the class-weighted learning strategy, and 2) word embedding feature on SMOTE-augmented outperform[1] all other classifiers [6]. On the other hand, the meta-classifier strategy is characterized by lower performance compared to the other strategies [6]. We refer to Table 2 in [6] for all details. In this study, we considered the models 1) SVM trained on not augmented data using class-weighted learning on bag-of-words (F1=0.48, precision=0.5, recall=0.45), 2) SVM trained on SMOTE-augmented data and word embeddings (F1=0.44, precision=0.35, recall=0.59) [6] as baseline.

V. RESULTS AND DISCUSSION

In this section, we discuss the main results of the experiments we performed to test the hypotheses from Section IV-A.

Hypothesis H1. In absence of data augmentation (and other strategies to cope with class imbalance, such as class-weighted learning), our results showed that the performance of the random forest classifier trained on the families of

[1]The combinations of learning strategy, NLP feature and classifier are evaluated on test data. The performance measures considered in study [6] are the area under curve (AUC), average precision (AP), precision, recall, F1-score and specificity.

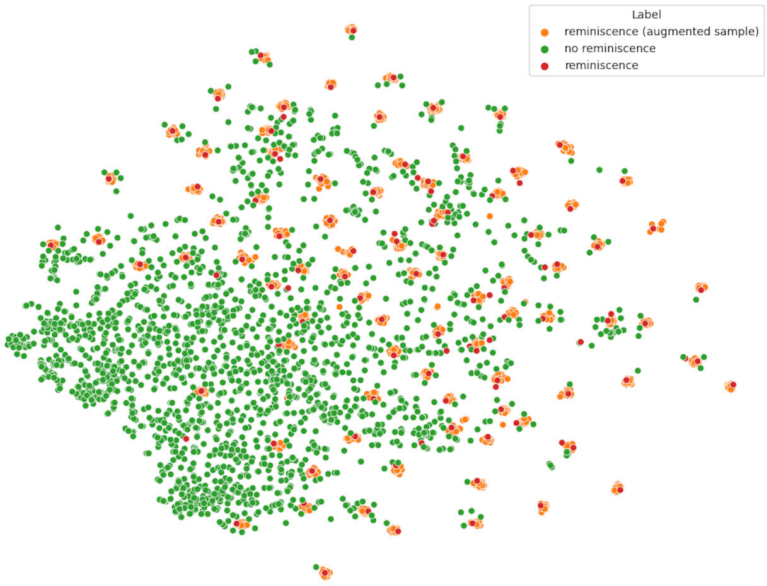


FIGURE 2. Visualisation of the t-SNE embeddings of transcripts with an augmentation factor $f = 20$. The samples generated through data augmentation cluster around the corresponding original transcripts.

TABLE 3. Top five performing combinations of features when training the random forest classifier with no data augmentation.

Features	Precision	Recall	F1
PTL_4, AMT_ICH, TEMP_WORDS, AMT_PT	0.466	0.145	0.216
DEP_AUX_PP, AMT_ICH, TEMP_WORDS, AMT_PT	0.474	0.127	0.197
PTL_4, AMT_ICH, TEMP_WORDS	0.438	0.127	0.191
TEMP_WORDS, AMT_PT	0.468	0.118	0.183
PTL_4, AMT_ICH, TIME, AMT_PT	0.405	0.118	0.182

features from Section III-B is quite poor. In Table 2 we show the four best performing features. We note that recall is worse than precision for all cases.

All other features are characterized by precision and recall equal to 0. This is equivalent to state that, in these cases, the classifier (trained without using data augmentation) does not predict correctly any of the 22 reminiscence samples in the test datasets. These sets are characterized by class imbalance, with 5% of samples labelled as reminiscence. On the other hand, when considering combinations of features, the performance of the random forest classifier improves, although without data augmentation. In Table 3 we show the best five combinations of features, in the case of no data augmentation.

It is interesting to note that the five best performing feature combinations comprise only four, three and two different families of features. These combinations comprise features from Table 3, among others. Therefore, it follows that the strategy of combining features to detect reminiscence with machine learning models delivers an increase of the F1-score with respect to the case where only a family of features is used (and with no data augmentation). For example, by combining **AMT_PT** features with those from the **PTL_4**, **AMT_ICH**, and **TEMP_WORDS** families, the classifier reaches F1=0.216, as opposed to F1=0.073, as shown in Table 2. Combining **DEP_AUX_PP** features with **AMT_ICH**, **TEMP_WORDS**, **AMT_PT**, the improved F1 score is almost seven times higher that the original one, instead (F1=0.197 as opposed to F1=0.034). In general, the combination of features allows improving both the precision and the recall of the classifier, with respect to the use of a single family of features. The improvement in recall is higher than the one in precision. By combining **AMT_PT** features with those from the **PTL_4**, **AMT_ICH**, and **TEMP_WORDS** families, the resulting recall is three times the one obtained by using **AMT_PT** features alone. However, the random forest classifier trained on the combinations of features in Table 3 (and without data augmentation) does not deliver better performance than the baseline models described in Section IV-D.

TABLE 4. Best features and data augmentation factors.

Features	Aug. factor f	Precision	Recall	F1
AMT_PT	4	0.296	0.564	0.388
PTL_4	4	0.317	0.564	0.404
DEP_AUX_PP	4	0.378	0.291	0.319
POS	12	0.293	0.555	0.383
BOW	14	0.490	0.455	0.466
EMB	16	0.325	0.664	0.436

TABLE 5. Best combination of features and corresponding augmentation factor.

Features	Aug. factor f	Precision	Recall	F1
LOC, PER, ORG, PTL_4, POS, AMT_ICH, TEMP_WORDS, EMB	6	0.535	0.509	0.519
LOC, PER, ORG, PTL_4, DEP_AUX_PP, POS, AMT_ICH, TIME, EMB	6	0.531	0.518	0.519
LOC, PER, ORG, PTL_4, DEP_AUX_PP, POS, AMT_ICH, AMT_PT, EMB	6	0.519	0.518	0.515
LOC, PER, ORG, PTL_4, POS, AMT_ICH, TIME, AMT_PT, EMB	8	0.463	0.600	0.521
LOC, PER, ORG, DEP_AUX_PP, POS, AMT_ICH, TIME, AMT_PT, EMB	8	0.459	0.591	0.515
LOC, PER, ORG, PTL_4, POS, TIME, TEMP_WORDS, AMT_PT, EMB	8	0.468	0.582	0.514

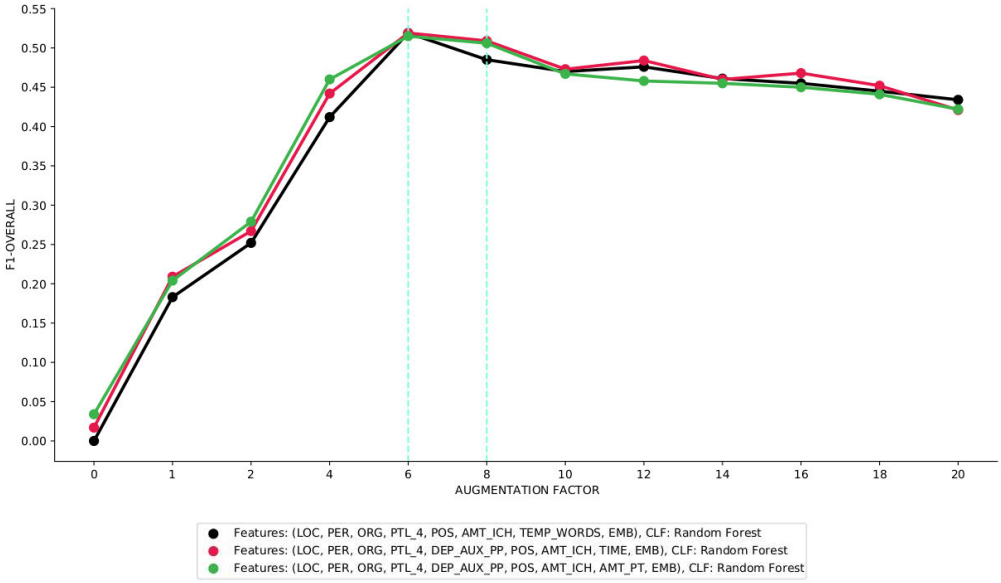


FIGURE 3. A plot representing the three feature combinations in the top half of Table 5.The vertical lines represent the augmentation factors where the best performance was achieved.

Hypothesis H2. The results of the conducted experiments show that data augmentation can lead to an improvement of performance in detecting reminiscence, as opposed to the case of no data augmentation. We start by considering data augmentation in the case of a single family of features, as shown in Table 4.

A small data augmentation factor (i.e., $f = 4$) allows achieving an F1-score equal to 0.388 for **AMT_PT**, as opposed to F1=0.073 in the case of no augmentation (see Table 2). Similarly, training the random forest classfier on **PTL_4** one reaches F1=0.404 as opposed to F1=0.033 when no augmentation is implemented. Finally, by training the classifier on **DEP_AUX_PP** one arrives at F1=0.319. The improvement in F1-score is due to an increase of both precision and recall. Higher augmentation factors, i.e., $f = 12, 14, 16$, allow for reaching F1=0.383 in the case of **POS** features, with F1=0.035 in the case of no augmentation (see Table 2). With an augmentation factor $f = 14$, **BOW** reaches F1=0.446 and word embeddings **EMB** achieve F1=0.436. Similar patterns are followed by all other features. In particular, time-related features such as **TIME** and **TIME_WORDS** reach a maximum F1-score for high augmentation factors, i.e., $f = 16$ and $f = 18$ respectively. We note that the random forest classifier trained on the (single families of) features in Table 4 and with data augmentation does not deliver better performance than the baseline models described in Section IV-D.

Table 5 shows the combinations of features and augmentation factors delivering highest performance. All combinations reach an F1-score above 0.51; they are characterized by different precision-recall trade-offs. The top half of Table 5 shows the combinations with highest precision; in the bottom half we report the combinations with highest recall, instead. We note that all combinations show an augmentation factor $f = 6$ or $f = 8$. All combinations comprise **POS**, **EMB** and named entity features, i.e., **LOC**, **PER**, **ORG**. In addition, the combinations with top precision comprise **PTL_4**, **AMT_ICH**, features. On the other hand, those with highest recall comprise **TIME**, **AMT_PT**. Interestingly, none of the combinations show bag-of-words, i.e. **BOW**.

Best models and comparison with the baseline. Table 5 collects the best performing models, i.e., random forest classifiers, given all combinations of features and data augmentation factors considered in this study. We show the top three best performing feature combinations in Figure 3, for all augmentation factors. We note that for all those feature combinations the performance of the random forest classifier increases by increasing the augmentation factor f, with a maximum corresponding to either $f = 6$ or $f = 8$ (see Table 5). Moreover Figure 3 shows that the performance of

the models slightly decreases for higher values of f. Table 5 shows that the best performing random forest classifier with highest precision is trained on an augmented dataset with factor $f = 6$. The model reaches an F1-score equal to 0.519, with precision and recall equal to 0.535 and 0.509, respectively. It delivers an improvement in performance when compared to the best baseline model, i.e., the SVM trained with class-weighted learning on bag-of-words and with no data augmentation [6] (see Section IV-D). In fact, the improvement in F1-score is equal to +8% and it is equal to +7% for precision and to +13% for recall. The improvement with respect the second baseline model, i.e., the SVM trained on word embeddings and using SMOTE-augmented data [6] (see Section IV-D) is equal to +18% for the F1-score and to +53% for precision. However, recall shows a decrease equal to -14%; it reaches a value equal to 0.509.

Table 5 shows that the best performing random forest classifier with highest recall is trained on an augmented dataset with factor $f = 8$. The model reaches an F1-score equal to 0.521, with precision and recall equal to 0.463 and 0.600, respectively. It improves the performance of both baseline models (see Section IV-D). In fact, considering the first baseline model, the improvement in F1-score is equal to +9% and to +33% for recall. However, precision shows a decrease equal to -7%. The improvement with respect to the second baseline model is equal to +18% for the F1-score, to +32% for precision and to +2% for recall.

VI. CONCLUSION AND FUTURE WORK

In this study we showed by an empirical evaluation that it is possible to improve the performance of reminiscence detection algorithms using data augmentation and the combination of different types of features. In our experiments we used only a simple technique, i.e., synonym replacement, to create new samples. However, at the core of the augmentation procedure proposed in our study we leveraged a pretrained BERT language model for the German language. We also generated different families of features from the corpus of transcripts and evaluated their combinations in multiple experiments. The results of our study strongly outperformed the baseline models that used SMOTE-augmented data or class weighted learning strategies on the same dataset from previous studies. Our best models achieved an F1-score equal to 0.52 on different combinations of features comprising, in particular, named entities, part-of-speech tags and word embeddings, and in presence of a mild data augmentation. The models show different precision-recall trade-offs and they considerably improved the precision and recall of the baseline models. As a consequence of the performance improvement with respect to the baseline models, the results of the proposed experiments may support the design of intelligent systems for the real-time detection of reminiscence of older adults in everyday settings. Therefore, the results of this study are significant for researchers in digital health, life span and aging research, as they would allow for the support of the well-being of older adults, by means of digital health interventions relying on the technology for the real-time detection of reminiscence in their lives. Moreover, the methods in this study can be straightforwardly applied to data collected in longitudinal studies and are not computationally-intensive. Despite the fact that our results do outperform the baseline models, there are still points which should be considered for future investigations. First, we did not perform any fine-tuning of the BERT embeddings during data augmentation. Second, for the feature extraction part of our study we relied on `spacy` and its pretrained models for POS-tagging, named entity recognition, sentence parsing and word embeddings for the German language. Moreover, the set of features used to detect reminiscence can be further extended, for example, by the use of Latent Dirichlet Allocation as in Yordanova *et al.*'s study [21]. Third, our main experiments were performed using only one machine learning classifier, i.e., random forests. Fourth, beyond synonym replacement, one could use other EDA techniques, such as random insertion random deletion, back-translation or the use of hyperonym substitution, as in the work of Yordanova *et al.* [21]. Finally, our experiments were conducted on data from a single naturalistic observation study. We will tackle these limitations in future work.

AUTHOR CONTRIBUTIONS

AF proposed the original line of research. TS designed and evaluated the natural language processing (NLP) and machine learning pipelines. AF and KY provided important intellectual inputs on the NLP and machine learning pipelines. KY prepared the first draft of the manuscript. TS, AF and KY finalized the manuscript. BD, ML and MM provided important intellectual inputs on reminiscence and the naturalistic observation study. All authors reviewed the finalized manuscript.

REFERENCES

[1] B. Demiray, M. Mischler, and M. Martin, "Reminiscence in everyday conversations: A naturalistic observation study of older adults," *J. Gerontol. B, Psychol. Sci. Social Sci.*, vol. 74, no. 5, pp. 745–755, Jun. 2019.

[2] S. Bluck and L. J. Levine, "Reminiscence as autobiographical memory: A catalyst for reminiscence theory development," *Ageing Soc.*, vol. 18, no. 2, pp. 185–208, Mar. 1998.

[3] G. J. Westerhof, E. Bohlmeijer, and J. D. Webster, "Reminiscence and mental health: A review of recent progress in theory, research and interventions," *Ageing Soc.*, vol. 30, no. 4, p. 697, 2010.

[4] S. Bluck, N. Alea, and B. Demiray, "You get what you need: The psychosocial functions of remembering," in *The Act of Remembering: Toward an Understanding of How We Recall the Past*, J. H. Mace, Ed. Hoboken, NJ, USA: Wiley, 2010, pp. 284–307.

[5] R. N. Butler, "The life review: An interpretation of reminiscence in the aged," *Psychiatry*, vol. 26, no. 1, pp. 65–76, Feb. 1963.

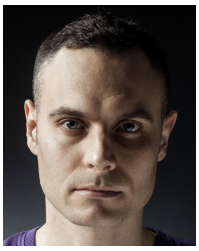
[6] A. Ferrario, B. Demiray, K. Yordanova, M. Luo, and M. Martin, "Social reminiscence in older adults' everyday conversations: Automated detection using natural language processing and machine learning," *J. Med. Internet Res.*, vol. 22, no. 9, Sep. 2020, Art. no. e19133.

[7] B. Demiray, M. Luo, A. Tejeda-Padron, and M. R. Mehl, "Sounds of healthy aging: Assessing everyday social and cognitive activity from ecologically sampled ambient audio data," in *Personality and Healthy Aging in Adulthood*. Cham, Switzerland: Springer, 2020, pp. 111–132.

[8] J. Leu, "Past places and social reminiscence," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., ACM Int. Symp. Wearable Comput.*, 2015, pp. 465–470.

[9] S. T. Peesapati, V. Schwanda, J. Schultz, M. Lepage, S.-Y. Jeong, and D. Cosley, "Pensieve: Supporting everyday reminiscence," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2010, pp. 2027–2036.
[10] P. E. Agroudy, T. Machulla, R. Rzayev, T. Dingler, M. Funk, A. Schmidt, G. Ward, and S. Clinch, "Impact of reviewing lifelogging photos on recalling episodic memories," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, Sep. 2016, pp. 1014–1019.
[11] N. T. Ly, J. Preßler, D. Gall, J. Hurtienne, and S. Huber, "Tangible interaction drawers for people with dementia: Retrieving living experiences from past memories," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, Sep. 2016, pp. 157–160.
[12] R. Barthel, M. de Jode, and A. Hudson-Smith, "Approaches to interacting with digital object memories in the real world," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 1179–1182.
[13] I. Li, A. K. Dey, and J. Forlizzi, "Understanding my data, myself: Supporting self-reflection with Ubicomp technologies," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 405–414.
[14] N. O'Rourke, D. B. King, and P. Cappeliez, "Reminiscence functions over time: Consistency of self functions and variation of prosocial functions," *Memory*, vol. 25, no. 3, pp. 403–411, Mar. 2017.
[15] J. D. Webster and X. Ma, "A balanced time perspective in adulthood: Well-being and developmental effects," *Can. J. Aging/La Revue Canadienne du Vieillissement*, vol. 32, no. 4, pp. 433–442, Dec. 2013.
[16] B. Funk, S. Sadeh-Sharvit, E. E. Fitzsimmons-Craft, M. T. Trockel, G. E. Monterubio, N. J. Goel, K. N. Balantekin, D. M. Eichen, R. E. Flatt, M.-L. Firebaugh, C. Jacobi, A. K. Graham, M. Hoogendoorn, D. E. Wilfley, and C. B. Taylor, "A framework for applying natural language processing in digital health interventions," *J. Med. Internet Res.*, vol. 22, no. 2, Feb. 2020, Art. no. e13855.
[17] P. Cappeliez and N. O'Rourke, "Empirical validation of a model of reminiscence and health in later life," *J. Gerontol. B, Psychol. Sci. Social Sci.*, vol. 61, no. 4, pp. P237–P244, Jul. 2006.
[18] N. O'Rourke, P. Cappeliez, and A. Claxton, "Functions of reminiscence and the psychological well-being of young-old and older adults over time," *Aging Mental Health*, vol. 15, no. 2, pp. 272–281, Mar. 2011.
[19] R. S. Gardner, A. T. Vogel, M. Mainetti, and G. A. Ascoli, "Quantitative measurements of autobiographical memory content," *PLoS ONE*, vol. 7, no. 9, Sep. 2012, Art. no. e44809.
[20] M. R. Mehl, "The electronically activated recorder (EAR): A method for the naturalistic observation of daily social behavior," *Current Directions Psychol. Sci.*, vol. 26, no. 2, pp. 184–190, Apr. 2017.
[21] K. Y. Yordanova, B. Demiray, M. R. Mehl, and M. Martin, "Automatic detection of everyday social behaviours and environments from verbatim transcripts of daily conversations," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Kyoto, Japan, Mar. 2019, pp. 242–251.
[22] *World Report on Ageing and Health*, World Health Org., Geneva, Switzerland, 2015.
[23] A. Lazar, H. Thompson, and G. Demiris, "A systematic review of the use of technology for reminiscence therapy," *Health Educ. Behav.*, vol. 41, no. 1, pp. 51S–61S, 2014.
[24] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*, vol. 10. Berlin, Germany: Springer, 2018, pp. 973–978.
[25] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[27] P. S. Bayerl and K. I. Paul, "What determines inter-coder agreement in manual annotations? A meta-analytic investigation," *Comput. Linguistics*, vol. 37, no. 4, pp. 699–725, Dec. 2011.
[28] L. Breiman, "Random forests," UC Berkeley, Berkeley, CA, USA, Tech. Rep. TR567, 1999.
[29] K. Crowston, X. Liu, and E. E. Allen, "Machine learning and rule-based automated coding of qualitative data," in *Proc. ASIS T Annu. Meeting Navigat. Streams Inf. Ecosyst.* Silver Springs, MD, USA: American Society for Information Science, 2010, pp. 108:1–108:2.
[30] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 2019, *arXiv:1901.11196*. [Online]. Available: http://arxiv.org/abs/1901.11196
[31] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 649–657.
[32] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," 2018, *arXiv:1805.06201*. [Online]. Available: http://arxiv.org/abs/1805.06201
[33] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[35] V. Marivate and T. Sefara, "Improving short text classification through global augmentation methods," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*. Cham, Switzerland: Springer, Aug. 2020, pp. 385–399.
[36] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional BERT contextual augmentation," in *Proc. Int. Conf. Comput. Sci.* Cham, Switzerland: Springer, Jun. 2019, pp. 84–95.
[37] V. Atliha and D. Šešok, "Text augmentation using BERT for image captioning," *Appl. Sci.*, vol. 10, no. 17, p. 5978, Aug. 2020.
[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[39] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," 2015, *arXiv:1511.06709*. [Online]. Available: http://arxiv.org/abs/1511.06709
[40] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, 1954.
[41] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Invest.*, vol. 30, no. 1, pp. 3–26, Jan. 2007.
[42] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith, "The TIGER treebank," in *Proc. Workshop Treebanks Linguistic Theories*, vol. 168, 2002, pp. 24–41.
[43] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2001.
[44] M. Luo, M. Neysari, G. Schneider, M. Martin, and B. Demiray, "Linear and nonlinear age trajectories of language use: A laboratory observation study of couples' conflict conversations," *J. Gerontol. B*, vol. 75, no. 9, pp. e206–e214, Oct. 2020.
[45] E. Ma. (Aug. 2019). *NLP Augmentation*. [Online]. Available: https://github.com/makcedward/nlp
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[47] M. Honnibal, I. Montani, S. V. Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," 2020. [Online]. Available: https://www.spacy.io
[48] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2002, pp. 857–864.
[49] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

TEODOR STOEV received the B.Sc. and M.Sc. degrees in computer science from the University of Rostock, Germany, where he is currently pursuing the Ph.D. degree. His research interests include natural language processing, machine learning, data fusion, behavioral models, and ontologies. From 2017 to 2020, he was a Student Trainee and a Research Assistant at one of the biggest German direct banks—the Comdirect Bank. In May 2020, he started working as a Researcher with the Junior Research Group, University of Rostock "Cognitive Methods for Situation-Aware Assistive Systems" (led by Dr.-Ing. Kristina Yordanova). He was also the Co-Chair of the 5th ARDUOUS Workshop 2021 on data annotation, which is affiliated with PerCom.



ANDREA FERRARIO received the Ph.D. degree in mathematics from ETH Zurich, in 2012. He has worked as a Consultant in analytics and AI for six years before his return to ETH, in 2018. Since then he has held a postdoctoral position as the Chair of Technology Marketing and the Scientific Director for the Mobiliar Lab for Analytics at ETH. His research interests include the intersection between philosophy and the applications of technology, with a focus on AI and mixed reality. His interests comprise the ethics and epistemology of AI, the use of natural language processing and machine learning for digital health interventions, and the use of immersive augmented reality to solve problems on the interpretability of machine learning models collaboratively.

BURCU DEMIRAY received the Ph.D. degree in developmental psychology from the University of Florida, in 2010. She worked as a Postdoctoral Researcher with the Psychology Institute, University of Zurich, in 2013. She is currently a Senior Researcher, a Research Group Leader, and the Gerontopsychology and Gerontology Chair with the Psychology Institute, University of Zurich. Her research interest includes real-life cognitive activities in the context of healthy longevity. She has an interdisciplinary research program in which she has been involved in and led various research projects with engineers, data scientists, and AI experts. She has taken active roles in the development and use of innovative research methods and devices (e.g., smartphone sensing and wearables) and state-of-the-art data analytics (e.g., machine learning) for the understanding and promotion of healthy longevity.

MINXIA LUO received the MSS degree in social science (gerontology) from The University of Hong Kong, Hong Kong, in 2014, and the Ph.D. degree in psychology (gerontopsychology) from the University of Zurich, Zurich, Switzerland, in 2020. She is currently a Postdoctoral Researcher with the University Research Priority Program "Dynamics of Healthy Aging," University of Zurich. Her research interests include daily activities in the context of cognitive health and well-being in older age. She works on mobile sensing data (e.g., audio recordings and accelerometers) and population surveys, ranging from seconds to decades.

MIKE MARTIN received the M.A. degree from the University of Georgia, USA, in 1990, the Ph.D. degree in psychology from the University of Mainz, in 1994, and the Habilitation degree from the University of Heidelberg, in 2001. Since 2002, he has been a Professor of gerontopsychology and gerontology with the University of Zurich and directing the Gerontology Center and the University Research Priority Program "Dynamics of Healthy Aging." He is an Honorary Professor with the University of Queensland. His research interests include the causes of healthy aging (WHO, 2015, 2020) and quality of life stabilization in real life. His group uses longitudinal studies combining measurements of abilities, skills, traits, goals, beliefs, high-density real-life activities, and environmental opportunities to test individual-specific and situation-aware predictor models of healthy aging for public health applications.

KRISTINA YORDANOVA received the bachelor's degree in computer engineering from the University of Duisburg-Essen, Germany, in 2008, the master's degree in artificial intelligence from Maastricht University, The Netherlands, in 2009, and the Ph.D. degree in ubiquitous computing from the University of Rostock, Germany, in 2014. She is currently the Head of the Junior Research Group "Cognitive Methods for Situation-Aware Assistive Systems" with the University of Rostock and a Research Associate with the University of Bristol, U.K. Her research interests include natural language processing, machine learning, and symbolic and probabilistic modeling with applications in assistive systems, social sciences, healthcare, and medicine.